NPS-AM-13-109

ACQUISITION RESEARCH PROGRAM SPONSORED REPORT SERIES

Lexical Link Analysis Application: Improving Web Service to Acquisition Visibility Portal

30 September 2013

Dr. Ying Zhao, Research Associate Professor,

Dr. Shelley P. Gallup, Research Associate Professor, and

Dr. Douglas J. MacKinnon, Research Associate Professor

Graduate School of Operational & Information Sciences

Naval Postgraduate School

Approved for public release; distribution is unlimited.

Prepared for the Naval Postgraduate School, Monterey, CA 93943.

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
30 SEP 2013		**00-00-2013 to 00-00-2013**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Lexical Link Analysis Application: Improving Web Service to Acquisition Visibility Portal	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Postgraduate School,Information Sciences Department,Monterey,CA,93943**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Lexical Link Analysis (LLA) is a form of text mining in which word meanings are represented in lexical terms (e.g., word pairs) of a word network. In the past, we have shown how LLA can systematically and automatically discover new patterns in large-scale defense acquisition data of multiple programs as indicators for program or investment performances. We also started to apply LLA to understand the quality of the data by comparing categories of information, detecting data and gaps. Last year, we examined the Acquisition Visibility Portal (AVP), which is a critical tool that provides the DoD-wide acquisition community with authoritative and accurate data services. We reported the first program from AVP to have undergone a relatively comprehensive LLA analysis. This year, we found that there is much consensus or consistency in the various categories (e.g., acquisition and engineering communities) of artifacts, yet gaps or low correlations seem to characterize the majority of the examined data for the relations among these categories. LLA, however, is able to discover in detail where the gaps and inconsistencies of the data reside. The depicted findings offered in this report can help decision-makers improve their resource and big data management, to better understand how particular acquisition strategies may affect the desired return on investment (ROI) among projects.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**43**	

The research presented in this report was supported by the Acquisition Research Program of the Graduate School of Business & Public Policy at the Naval Postgraduate School.

To request defense acquisition research, to become a research sponsor, or to print additional copies of reports, please contact any of the staff listed on the Acquisition Research Program website (www.acquisitionresearch.net).

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL

Abstract

Lexical Link Analysis (LLA) is a form of text mining in which word meanings are represented in lexical terms (e.g., word pairs) of a word network. In the past, we have shown how LLA can systematically and automatically discover new patterns in large-scale defense acquisition data of multiple programs as indicators for program or investment performances. We also started to apply LLA to understand the quality of the data by comparing categories of information, detecting data and gaps. Last year, we examined the Acquisition Visibility Portal (AVP), which is a critical tool that provides the DoD-wide acquisition community with authoritative and accurate data services. We reported the first program from AVP to have undergone a relatively comprehensive LLA analysis. This year, we found that there is much consensus or consistency in the various categories (e.g., acquisition and engineering communities) of artifacts, yet gaps or low correlations seem to characterize the majority of the examined data for the relations among these categories. LLA, however, is able to discover in detail where the gaps and inconsistencies of the data reside. The depicted findings offered in this report can help decision-makers improve their resource and big data management, to better understand how particular acquisition strategies may affect the desired return on investment (ROI) among projects.

Keywords: lexical link analysis, text mining, acquisition visibility portal, unstructured data, data quality, authoritative data service, accurate data service, resource management

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgments

We thank Mr. Robert Flowe from AT&L/ARA/FP&OS, OSD, who provided sponsorship to access the Acquisition Visibility Portal and relevant questions along with insightful discussions.

THIS PAGE INTENTIONALLY LEFT BLANK

About the Authors

Dr. Ying Zhao is a research associate professor at the Naval Postgraduate School (NPS). Dr. Zhao joined NPS in May 2009. Her research is focused on knowledge management approaches such as data/text mining, Lexical Link Analysis, search and visualization for system self-awareness, decision-making, and collaboration. She received her PhD in mathematics from MIT and co-founded Quantum Intelligence, Inc. She was principal investigator (PI) for six contracts awarded by the DoD Small Business Innovation Research (SBIR) Program. She was the co-author of two U.S. patents in knowledge pattern search from networked agents and in fusion and visualization for multiple anomaly detection systems.

Dr. Ying Zhao
Information Sciences Department
Naval Postgraduate School
Monterey, CA 93943-5000
Tel: 831-656-3789
Fax: (831) 656-3679
E-mail: yzhao@nps.edu

Dr. Shelley Gallup is a research associate professor at the Naval Postgraduate School's Department of Information Sciences, and the director of Distributed Information and Systems Experimentation (DISE). Dr. Gallup has a multidisciplinary science, engineering, and analysis background, including microbiology, biochemistry, space systems, international relations, strategy and policy, and systems analysis. He returned to academia after retiring from naval service in 1994 and received his PhD in engineering management from Old Dominion University in 1998. Dr. Gallup joined NPS in 1999, bringing his background in systems analysis, naval operations, military systems, and experimental methods first to the Fleet Battle Experiment series (1999–2002) and then to the FORCEnet experimentation in the Trident Warrior series (2003–present).

Dr. Shelley P. Gallup
Information Sciences Department
Naval Postgraduate School
Monterey, CA 93943-5000
Tel: 831-656-1040
Fax: (831) 656-3679
E-mail: spgallup@nps.edu

Dr. Doug MacKinnon is a research associate professor at the Naval Postgraduate School (NPS). Dr. MacKinnon is the deputy director of the Distributed Information and Systems Experimentation (DISE) research group where he leads multi-

disciplinary studies ranging from maritime domain awareness (MDA), to knowledge management (KM) and Lexical Link Analysis (LLA). He also led the assessment for the Tasking, Planning, Exploitation, and Dissemination (TPED) process during the Empire Challenge 2008 and 2009 (EC08/09) field experiments and for numerous other field experiments of new technologies during Trident Warrior 2012 (TW12). He teaches courses in operations research. He holds a PhD from Stanford University, conducting successful theoretic and field research in Knowledge Management (KM). He has served as the program manager for two major government projects of over $50 million each, implementing new technologies while reducing manpower requirements. He has served over 20 years as a naval surface warfare officer, amassing over eight years at sea and serving in four U.S. Navy warships with five major, underway deployments.

Dr. Douglas J. MacKinnon
Information Sciences Department and Graduate School of Operational and Information Sciences
Naval Postgraduate School
Monterey, CA 93943-5000
Tel: 831-656-1005
Fax: (831) 656-3679
E-mail: djmackin@nps.edu

NPS-AM-13-109

ACQUISITION RESEARCH PROGRAM SPONSORED REPORT SERIES

Lexical Link Analysis Application: Improving Web Service to Acquisition Visibility Portal

30 September 2013

Dr. Ying Zhao, Research Associate Professor,

Dr. Shelley P. Gallup, Research Associate Professor, and

Dr. Douglas J. MacKinnon, Research Associate Professor

Graduate School of Operational & Information Sciences

Naval Postgraduate School

Disclaimer: The views represented in this report are those of the author and do not reflect the official policy position of the Navy, the Department of Defense, or the federal government.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

Executive Summary

We define *awareness* as the cognitive interface between decision-makers and a complex system, expressed in a range of terms or features, or specific vocabulary or lexicon, to describe the attributes and surrounding environment of the system. Lexical Link Analysis (LLA) is a form of text mining in which word meanings represented in lexical terms (e.g., word pairs) can be represented as if they are in a community of a word network. In the past, we have explored how LLA systematically and automatically discovers new patterns that were previously unknown, and identifies data dependencies from large-scale defense acquisition data of multiple programs that might be indicators for program or investment performances in defense acquisition decision-making and research communities.

We also started to apply LLA to improve our understanding of the quality of the data by comparing categories of information and by detecting data overlaps, inconsistency, and gaps from a single program point of view. The Acquisition Visibility Portal (AVP) is a critical tool that provides the DoD-wide acquisition community with authoritative and accurate data services via interfaces to Defense Technical Center (DTIC) and Defense Acquisition Management Information Retrieval (DAMIR) for programs (e.g., major defense acquisition programs [MDAPs], acquisition category II [ACATII] programs) with milestones, costs, schedules and performance data, selected acquisition reports (SARs), acquisition strategy reports (ASRs), the systems engineering plans (SEPs), the test & evaluation master plans (TEMPs), and the defense acquisition executive summary (DAES), among others.

The major advantage of using LLA is to reveal and depict—to decision-makers—the correlations, associations, and program gap identifications across all the programs in the AVP over many years. This enables strategic understanding of data gaps and potential trends, and can inform managers what areas might be highly risky for a program and how resource and big data management might affect the desired return on investment (ROI) among projects .

We performed a relatively comprehensive LLA analysis to generate semantic networks developed from acquisition artifacts among multiple categories of program data. First, our effort revealed that there exist high data correlations among many areas in the various artifacts. Yet, gaps or low correlations seem to characterize the relations between these categories of artifacts, for example, between ASRs and SEPs & TEMPs; between SEPs and TEMPs; and between SARs and DAESs. Specifically, many concepts in one category are not documented in another, which could form the basis for further inquiry or future reconciliation of the expectations (e.g., acquisition strategy) and realities (e.g., engineering feasibility) from various

communities for the same MDAP program. LLA is able to discover in detail where the gaps and inconsistencies of the data across multiple categories of informations (e.g., ASRs, SEPs & TEMPs), help identify the issues, and offer specific and productive directions for further examination regarding why there are gaps and where they exist. These are outlined in the FOUO appendices of this report.

Lexical Link Analysis Application: Improving Web Service to Acquisition Visibility Portal

Background

It is critical that the Department of Defense (DoD)-wide acquisition community can access authoritative and accurate data services for decision-making. The Acquisition Visibility Portal (AVP) was such a data service that achieved this purpose by interfacing with program elements and warfighter requirements via a Defense Technical Information Center (DTIC) interface (program elements [PEs] [see http://www.dtic.mil/descriptivesum/] and requirements [see http:/www.dtic.mil/doctrine]). The AVP also included an interface to Defense Acquisition Management Information Retrieval (DAMIR; http://www.acq.osd.mil/damir/) to access large, detailed collections of information such as milestones, costs, schedules, and performance data of selected acquisition reports (SARs) and acquisition Strategy Reports (ASRs), among others, regarding Major Defense Acquisition Programs (MDAPs) and Acquisition Category II (ACATII) program data. The AVP provided automatic methodologies to systematically improve understanding of the quality of the data.

In the past, we have explored an analytic and visualization tool named Lexical Link Analysis (LLA), which we applied to various areas of acquisition research, for example, to link warfighter requirements with the acquisition programs and program elements (Gallup, MacKinnon, Zhao, Robey & Odell, 2009; Zhao, Gallup & MacKinnon, 2010,2011a,2011b,2011c,2011d,2012a,2012b,2013; Zhao, Brutzman & MacKinnon 2013).

Recently, we have started to explore how LLA can help detect data quality, inconsistency, gaps, or bad data among categories of data by automatically discovering new patterns that were previously unknown and by identifying data dependencies that might be indicators for program or investment performances.

For example, one of the biggest risk factors in defense acquisition is the unanticipated effects of program interactions. Some current work exists toward identifying interdependence among programs within a system of systems (SoS; Dahmann et al., 2005). Yet, more broadly, and as a result of required joint capabilities, portfolios often include program interdependencies and SoS effects. Ultimately, the current “program-centric” acquisition paradigm is increasingly ill-suited to identify and address program risks that arise outside of program boundaries. LLA can help isolate these issues from the body of information collected, but have yet to be effectively identified.

Furthermore, we also observed that very little of the information generated for program oversight is amenable to effective analysis. Every major acquisition program's milestone review generates volumes of information, which the Office of the Secretary of Defense (OSD) staff is supposed to review to determine if the program is properly prepared for the next milestone. Although acquisition professionals and decision-makers at OSD are beginning to compile these artifacts centrally to facilitate review and analysis, at present the only way to analyze the information in these artifacts is to read them. With limitations on staffing, little time is available to thoroughly review these artifacts. Moreover, each functional community is required to review only the particular document for which it is responsible. For example, the systems engineering community typically only examines the systems engineering plans (SEPs), the test and evaluation community looks only at the test & evaluation master plans (TEMPs), and the acquisition community looks at the acquisition strategy reports (ASRs). Rarely do any of these stakeholders review multiple reports or jointly discuss them to determine if they are mutually consistent and consider inconsistencies that might indicate programmatic risk. There is even less incentive and opportunity to look for external factors that would potentially invalidate the assumptions that underpin the basic cost, schedule, and performance targets of each program's execution.

Motivated by these situations, we applied LLA as one of automatic tools to examine large collections of artifacts for many programs in various categories across the acquisition and engineering communities. By using LLA, one can learn from the actual data to see how the common concepts are expressed in different artifacts and communities. Overlaying the concepts for each category of artifacts to conduct a pair-wise comparison exposes significant disconnections between them. The automatic discovery of the disconnection or gaps could be fed back to the human analysts or decision-makers to perform further investigations.

Methodology

Detect Data Gaps in the AVP Categories

To detect the data gaps between two categories of information, LLA compares these artifacts from one category to another, for example, comparing the ASRs with the SEPs at Milestone B. These comparisons, reported as themes, concepts, and word pairs, may help cue a decision-maker's attention to the potential issues and consider specific and productive directions for further scrutiny.

To illustrate the methodology, we first extracted a sample Navy ship-building program as a representative of Major Defense Acquisition Programs (MDAPs) from the AVP with categories of information in the following documents:

- SEPs: Systems Engineering Plans

- TEMPs: Test & Evaluation Master Plans
- ASRs: Acquisition Strategy Reports
- SARs: Selected Acquisition Reports
- DAESs: Defense Acquisition Executive Summaries
- Cert 2366b: Certification Milestone B, Acquisition Decision Memorandum (ADM)
- APB: Acquisition Program Baseline
- TRA: Technology Readiness Assessment

When using LLA to compare two categories of information, for example, comparing ASRs, SEPs/TEMPs, we asked the question, "What are the concepts or clusters of concepts discussed in ASRs but not discussed in SEPs and TEMP?" If found, there might be various reasons to explain the discrepancies. For example, if a cluster of concepts appear only in ASRs but not in SEPs and TEMPs, it could be a gap because of (1) a data quality issue (e.g., a mishandling of data by AVP), (2) a data classification issue (e.g., unclassified data vs. classified data), or (3) a real requirement gap (i.e., a concept required by acquisition for which no engineering feasibility document or blueprint can be located). These types of information, if detected earlier, would provide decision-makers with the basis to make earlier amendments, thereby reducing program risks and costs in the future.

In this report, we report the overall pair-wise comparisons of these categories. FOUO content is not included in the main body of the report. For official use only (FOUO) content is reported in Appendices A, B, C, and D:

- Appendix A—Compare ASRs and SEPs/TEMPs: Are there concepts or clusters of concepts discussed in ASRs but not discussed in SEPs and TEMPs?
- Appendix B—Compare SEPs and TEMPs: Although they are both generated in the engineering community, what are the gaps between the two?
- Appendix C—Compare SARs and DAES: These two categories should contain similar information, but one is unclassified (SARs) and the other is FOUO (DAESs); where are the gaps?
- Appendix D—Compare 2004 DAESs and 2010 DAESs: What are the new program components added to the program from one time point to another?

These questions can be answered by human analysts if they have enough time to go through the piles of artifacts one by one and mark the differences. The advantage of using LLA is to automate the process and provide initial screening for human analysts.

In the next section, we first review Lexical Link Analysis to provide a base for how a comparison is done.

Overview of Lexical Link Analysis

As in military operations, where the term *situational awareness* was coined, we note that that our efforts can inform *awareness* of analyzed data in a unique way that helps improve a decision-maker's understanding or awareness of the data's content. We, therefore, define awareness as the cognitive interface between decision-makers and a complex system, expressed in a range of terms or features, or a specific vocabulary or lexicon, to describe the attributes and surrounding environment of the system. Specifically, LLA is a form of text mining in which word meanings represented in lexical terms (e.g., word pairs) can be represented as if they are in a community of a word network.

Link analysis discovers and displays a network of word pairs. These word pair networks are characterized by one-, two-, or three-word themes. Figure 1 shows a visualization of common lexical links shared between Systems 1 and 2, shown in the red box. A system, or a corpus, can be a collection of documents for an actual physical system (e.g., acquisition strategies for a Navy ship-building program) or simply a category of information. A node in Figure 1 represents a word in a corpus and a link or edge represents a word pair. A word pair is a bi-gram (Manning & Schütze, 1999) word pair extracted from the corpus. Within the field of computational linguistics, an *n*-gram is a sequence of *n* items matched to certain probabilistic patterns from a given text. Size 2 of an *n*-gram is a bi-gram. In Figure 1, each link color refers to the collection of words, lexicon, or features that belongs to a cluster that describes a concept or theme. In overlapping areas, nodes are *lexically linked*. Unlinked, outer vectors (outside the red box) indicate unique system features. Figure 2 shows the information from three categories that can be compared, and Figure 3 shows the information from two time periods that can be compared. What is unique here is that LLA constructs these linkages via intelligent agent technology using social network grouping methods.

The closeness of the systems in comparison can be examined visually or using the quadratic assignment procedure (QAP; Hubert & Schultz, 1976 [e.g., in UCINET]; Borgatti, Everett, & Freeman, 2002) to compute the correlation of two sets of lexical terms from two systems and analyze the structural differences in the two systems, as shown in Figure 4.

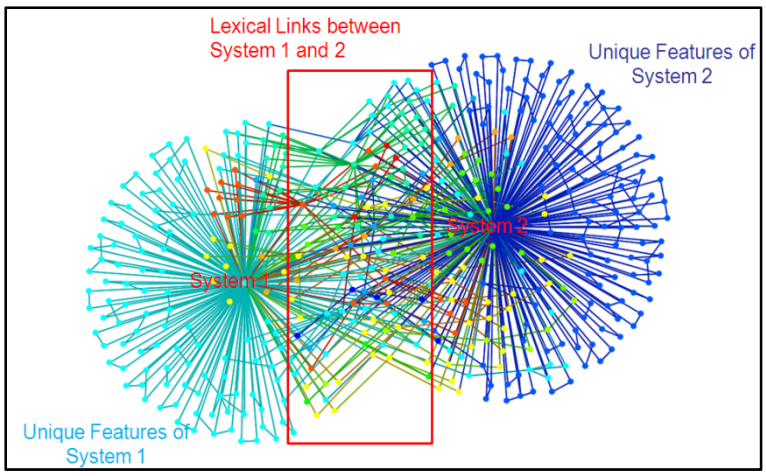


Figure 1. Comparing Two Systems Using LLA

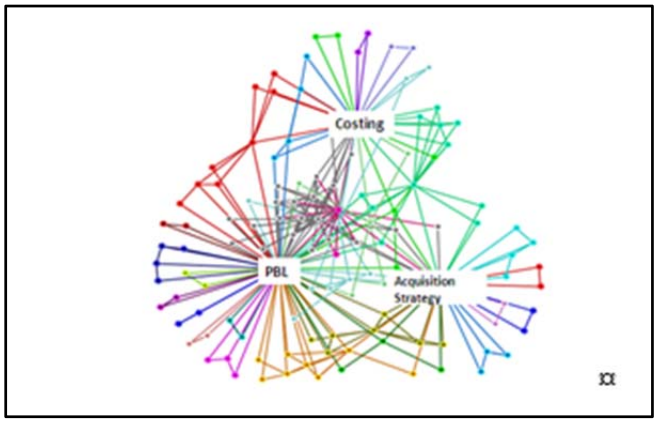


Figure 2. Comparing Three Categories

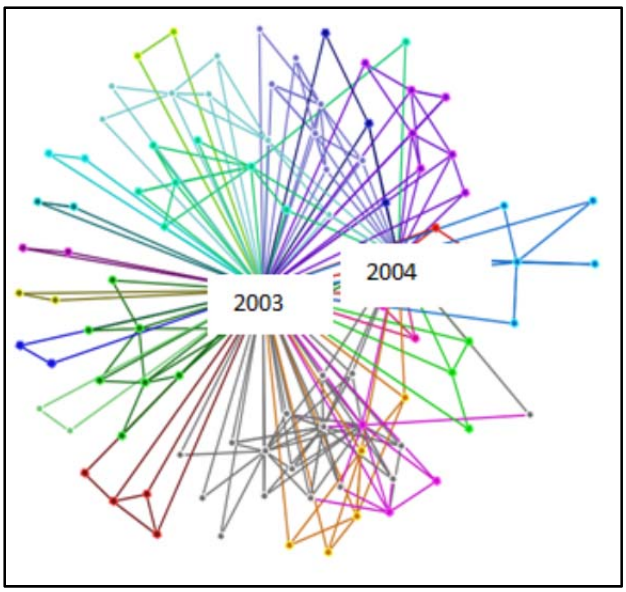


Figure 3. Comparing Two Time Periods

QAP Correlations

		1 lla_n	2 lla_n	3 lla_n	4 lla_n	5 lla_n	6 lla_n	7 lla_n	8 lla_n
1	lla_network_1_2010-AcquisitionStrategy	1.000	0.174	0.156	0.155	0.036	0.111	0.020	0.062
2	lla_network_1_2003-AcquisitionStrategy	0.174	1.000	0.447	0.149	0.052	0.119	0.043	0.089
3	lla_network_1_2004-AcquisitionStrategy	0.156	0.447	1.000	0.111	0.047	0.119	0.051	0.080
4	lla_network_1_2005-AcquisitionStrategy	0.155	0.149	0.111	1.000	0.156	0.084	0.034	0.088
5	lla_network_1_2006-AcquisitionStrategy	0.036	0.052	0.047	0.156	1.000	0.067	0.036	0.056
6	lla_network_1_2007-AcquisitionStrategy	0.111	0.119	0.119	0.084	0.067	1.000	0.097	0.123
7	lla_network_1_2008-AcquisitionStrategy	0.020	0.043	0.051	0.034	0.036	0.097	1.000	0.286
8	lla_network_1_2009-AcquisitionStrategy	0.062	0.089	0.080	0.088	0.056	0.123	0.286	1.000

QAP P-Values

		1 lla_n	2 lla_n	3 lla_n	4 lla_n	5 lla_n	6 lla_n	7 lla_n	8 lla_n
1	lla_network_1_2010-AcquisitionStrategy	0.000	0.020	0.020	0.020	0.020	0.020	0.020	0.020
2	lla_network_1_2003-AcquisitionStrategy	0.020	0.000	0.020	0.020	0.020	0.020	0.020	0.020
3	lla_network_1_2004-AcquisitionStrategy	0.020	0.020	0.000	0.020	0.020	0.020	0.020	0.020
4	lla_network_1_2005-AcquisitionStrategy	0.020	0.020	0.020	0.000	0.020	0.020	0.020	0.020
5	lla_network_1_2006-AcquisitionStrategy	0.020	0.020	0.020	0.020	0.000	0.020	0.020	0.020
6	lla_network_1_2007-AcquisitionStrategy	0.020	0.020	0.020	0.020	0.020	0.000	0.020	0.020
7	lla_network_1_2008-AcquisitionStrategy	0.020	0.020	0.020	0.020	0.020	0.020	0.000	0.020
8	lla_network_1_2009-AcquisitionStrategy	0.020	0.020	0.020	0.020	0.020	0.020	0.020	0.000

QAP statistics saved as datafile QAP Correlation Results

Figure 4. QAP Correlation via UCINET

Figure 5 shows a visualization of LLA with connected keywords or concepts as clusters, groups, or themes. Words are linked as word pairs that appear next to each other in the original documents. Different colors indicate different clusters of word groups. They were produced using a social network community detection method (Girvan & Newman, 2002) where words are connected, as shown in a single color, as if they are in a social community. A word center is formed around a word node connected with a list of other words in word pairs. For instance, Figure 6 shows a detailed view of a theme or word group in Figure 5. The center words are analysis, research, and approach. In this example, we use three words such as “analysis,

research, approach" to label such a group, where the top three words are those with the highest total degree of centralities (Freeman,1979; Wasserman & Faust, 1994).

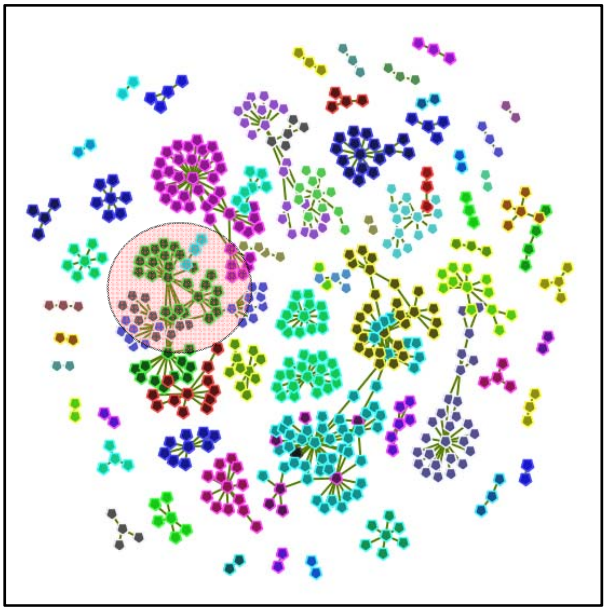


Figure 5. Word and Term of Themes Discovered and Shown in Colored Groups

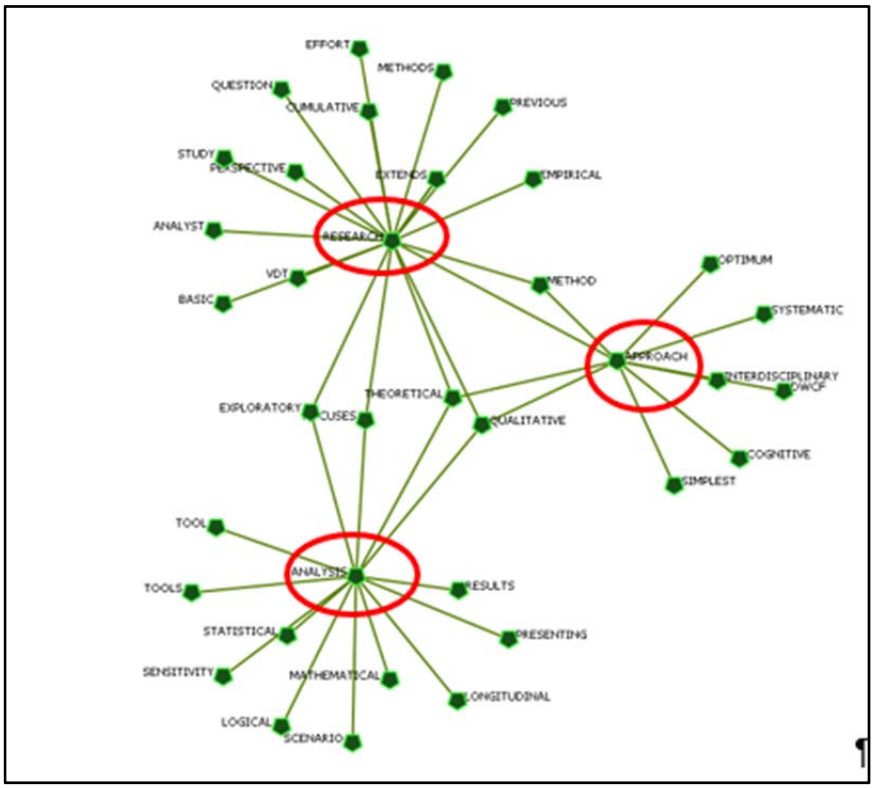


Figure 6. A Detailed View of a Theme or Word Group From Figure 5

The detailed steps of LLA processing include the following:

- Step 1: Select word pairs based on the following bi-gram parameters:
  - the probability threshold for one word next to another word in a word pair and
  - the minimum frequency for each individual word.
- Step 2: Apply a social network community-finding algorithm (i.e., Newman community detection method; Girvan & Newman, 2002) to group the word pairs into themes. A theme includes a cluster of lexical word pairs connected to each other.
- Step 3: Compute a "weight," or an importance measure, for a theme.
- Step 4: Sort theme weights by time and study the distributions of the themes by time.

The outputs of LLA include lexical network visualizations such as the ones in Figures 1, 2, 3, 4, 5, and 6; radar visualization; and matrix visualization (Zhao, Gallup, & MacKinnon, 2010). The word pair groups or themes as shown Figure 5 and 6 are further divided into three types according to the weights in Step 3:

- Popular (P): themes containing the highest number of mutually connected word pairs. The themes represent the main topics in a corpus at the time. The theme represented in Figure 6 is an example of a popular theme.
- Emerging (E): themes containing the medium number of mutually connected word pairs. These themes may grow to be popular over time.
- Anomalous (A): themes containing the lowest number of mutually connected word pairs. These themes may be off-topics compared to other themes and may be interesting for further investigation.

Business Problems That LLA Addresses

As a text analysis tool, LLA typically addresses the business problems of discovering themes and topics in unstructured documents and sorting the importance of the themes accordingly. Current methods, for example, internet search methods of ranking pages, require established hyperlinks, citation networks, or other forms of crowd-sourced collective intelligence. LLA is especially useful for data without hyperlinks and citation networks, for example, large-scale government internal documents. Furthermore, current methods typically rank the importance of the information based on its popularity. For example, we found that in many

business applications, it is useful to rank information based on emerging importance or anomalousness.

Current research on social network analysis focuses mostly on people or organizations with direct associations regardless of the contents linked. The so-called study of centrality (Girvan & Newman, 2002; Feldman & Sanger, 2007) has been a focal point for the social network structure study. Finding the centrality of a network lends insight into the various roles and groupings such as the connectors (e.g., mavens, leaders, bridges, isolated nodes), the clusters (and who is in them), the network core, and its periphery (Orgnet, 2011).

One of the core innovations of LLA is to analyze the content (e.g., documents and social media communications) created by social entities (e.g., people or organizations) and, therefore, create alternative networks (i.e., semantic networks) to traditional social networks. The resulting networks from LLA examine both social and semantic networks in terms of the organizations and people involved in the important themes, and how semantic networks might suggest improved potential collaborations and predict future outcomes.

Implementation Details

In the past year, we continued our efforts at the Naval Postgraduate School (NPS) by using collaborative learning agents (CLAs; QI, 2009) and other tools, including AutoMap (Center for Computational Analysis of Social and Organizational Systems [CASOS], 2009) for improved visualizations. Results from these efforts arose from leveraging intelligent agent technology via an educational license with Quantum Intelligence, Inc. CLA is a computer-based learning agent, or agent collaboration tool, capable of ingesting and processing data sources.

We have been generating visualizations including a lexical network visualization using various open source tools. We began by using the Organizational Risk Assessment (ORA; CASOS, 2009) tool and expanded to other tools. For example, in the past year, we developed 3-D network views using Pajek (Batagelj, Mrvar & Zaveršnik, 2011) and X3D (Web3D, 2011). We also developed our visualizations radar view and match view (Zhao at al., 2010).

LLA uses a computer-based learning agent called CLA (QI, 2009) to employ an unsupervised learning process that separates patterns and anomalies. Unsupervised agent learning is implemented by indexing each set of documents separately and in parallel using multiple learning agents. Unsupervised agents are used because the learning data for supervised agents are expensive to obtain. Multiple agents can work collaboratively and in parallel. We set up a cluster utilizing Linux servers in the NPS High Performance Computing Center (HPC) to handle the

large-scale data and the secure environment in the NPS Secure Technology Battle Laboratory (STBL).

Relations to Other Methods

The LLA approach is more properly related to latent semantic analysis (LSA; Dumais, Furnas, Landauer, & Deerwester, 1988) and probabilistic latent semantic analysis (PLSA). In the LSA approach, a term-document matrix is the starting point for analysis. The elements of the term-document or feature-object (term as feature and document as object) matrix are the occurrences of each word in a particular document (i.e., *A = [*a_{ij}*]*, where a_{ij} denotes the frequency in which term *j* occurs in document *i*). The term-document matrix is usually sparse. LSA uses singular value decomposition (SVD) to reduce the dimensionality of the term-document matrix. SVD cannot be applied to the cases where the vocabulary (the unique number of terms) in the document collection is large; for example, the number of unique terms in the DoD's acquisition documentation approaches the "large" value that would make SVD inapplicable. LSA has been widely used to improve information indexing, search/retrieval, and text categorization.

A recent development related to this method is called latent Dirichlet allocation (LDA; Blei, Ng, & Jordan, 2003), which is a generative probabilistic model of a corpus. In LDA, a document is considered to be composed of a collection of words, a "bag of words," where word order and grammar are not considered important. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a statistical distribution (*Dirichlet* distribution) over the corpus.

Our theme generation from LLA is different than from LDA, in which a collection of lexical terms is connected to each other semantically, as if the terms are in a social community, and social network grouping methods are used to group the words. Also unlike LSA, our method is easily scaled to analyze a large vocabulary and is generalizable to any sequential data.

Anticipated Benefits

Our LLA method provides solutions to meet the critical needs of the acquisition research community. The key advantage is to provide an innovative near-real-time self-awareness system to transfer diversified data services into strategic decision-making knowledge, specifically through the following:

- Automation: High correlation of LLA results—with the link analysis done by human analysts—makes it possible to save human power and improve responsiveness. Automation is achieved via computer

program or software agents to perform LLA frequently—and in near real-time.

- Discovery: LLA "discovers" and displays a network of word pairs. These word pair networks are characterized by one-, two-, or three-word themes. The weight of each theme is determined based on its frequency of occurrence. LLA may also discover blind spots of human analysis that are caused by the overwhelming amount of data for human analysts to consider.
- Validation: LLA may provide different perspectives of links. In the acquisition context, links discovered by human analysts may emphasize component and part connections that do not necessarily reflect content overlaps.Consequently, it can provide improved results in terms of trust, quality of association, and discovery; can help break through the taxonomy of ignorance (Denby & Gammack, 1999) and organizational boundaries; and can help improve organizational reach.

Research Results

Data Access Issues

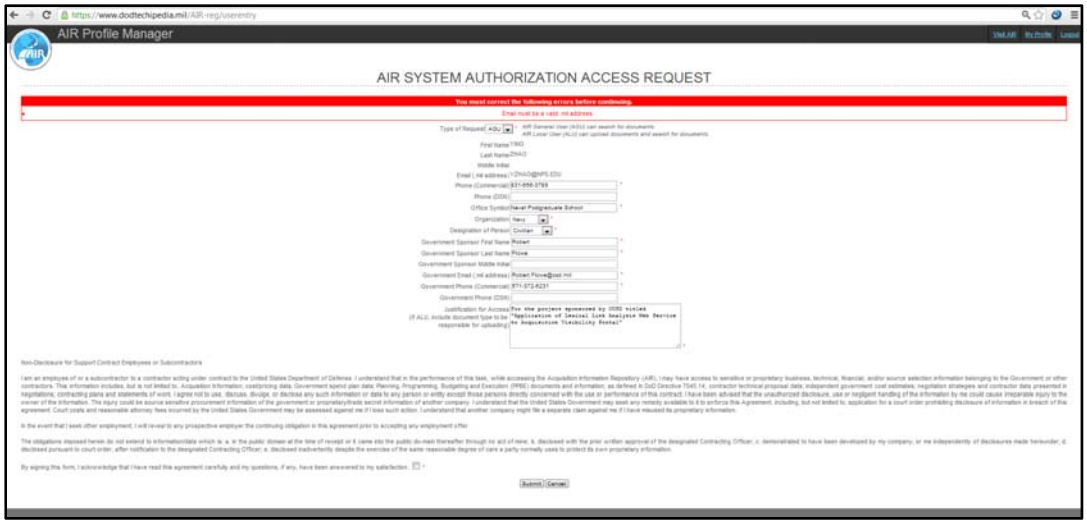
One of the issues we faced during this project was gaining access to the relevant repositories of authoritative acquisition data. As government researchers performing OSD-sponsored work, we hope that there will be no legitimate reason for restricting access to the following data sites in the future:

- Acquisition Visibility Portal (https://portal.acq.osd.mil/portal/server.pt/community/acquisition_visibility/1427)
- DAMIR (https://ebiz.acq.osd.mil/DAMIR/PortalMain/DamirPortal.aspx)
- Earned Value Management Central Repository (EVM-CR): https://service.dcarc.pae.osd.mil/EVM/

These three sites can typically be successfully accessed with a Common Access Card (CAC) login. We were able to view the lists of MDAPs (Major Defense Acquisition Programs) and MAISs (Major Automated Information Systems), but the actual data required for research reside mostly in the Acquisition Information Repository (AIR; https://www.dodtechipedia.mil/dodc/plugins/AIR/airdocuments.action)

The AIR site (Figure 7[a]) is where we had major access issues. The AIR profile manager requires a .mil email to allow access. As NPS research faculty, we have mail aliases with .mil. We tried to register in AIR using these .mil email aliases, yet the registration profile form did not allow us to change the email field, as shown in Figure 7(b). Later we found that the AIR system takes the email address from the CAC certificate and cannot be edited. We later also explored if we might add the .mil address to our CAC cards. We checked with the ID office and the NPS technology support center and were told that an NPS CAC card can only contain one email address. Although a switch can be made to the CAC card indicating a .mil address, it would impact our other functions. The documented difficulty seems related to the policy-related, procedural, administrative, or technical issues that prevent authorized NPS researchers from gaining access to authoritative acquisition data to conduct OSD-sponsored acquisition research.

(a)



(b)

Figure 7. Difficulty Accessing AIR

Because of these data access difficulties, initial sample data were extracted manually for this report to show the feasibility and importance of applying LLA. We

report the results in detail using manually extracted data for one Navy ship-building program in the following section.

Results

To develop comprehensive LLA comparisons, we first extracted a sample Navy ship-building program as a representative of MDAPs from the AVP with categories of information to demonstrate the method as follows:

- SEPs: Systems Engineering Plans, two documents, 222 pages
- TEMPs: Test & Evaluation Master Plan, five documents, 62 pages
- ASRs: Acquisition Strategy Report,11 documents including metrics, 634 pages
- SARs: Selected Acquisition Report, nine documents, 313 pages
- DAESs: Defense Acquisition Executive Summaries, 19 documents, 447 pages
- Cert 2366b: Certification Milestone B, Acquisition Decision Memorandum (ADM), 12 documents, 105 pages
- APBs: Acquisition Program Baseline, three documents, 39 pages
- TRA: Technology Readiness Assessment, one document, one page

For each comparison pair for two categories of information, we use the ratio of the number of word pairs that appear in both categories and the total number of word pairs as an overall correlation for each pair. For example, Figure 8 lists the top 20 themes discovered by comparing data for Acq_Str or the ASRs and SEPs or SEPs/TEMPs with the highest correlations. In Row 2, there are 299 word pairs for the two sources together classified in Theme 117(E), and 47 of them appear in both sources, indicating potential feature overlaps. The *correlation* is the ratio, which is 47/299 = .157. This indicates 15.7% of the features represented as word pairs were shared in both artifacts. As a detail shown in Figure 9, parts of the 299 word pairs in Theme 117(E) are visualized in red, yellow, and green links, representing the shared word pairs, pairs unique to the ASRs, and pairs in the SEPs/TEMPs, respectively. Figure 10 lists the least correlated themes discovered by comparing data for the ASRs and SEPs/ TEMPs. In Row 2, there are 149 word pairs for the two sources together, classified in Theme 359(E)(A), and four of them appear in both sources (overlap). The correlation is the ratio, which is 4/149 = .027. The detail shown in Figure 11, parts of the 149 word pairs in Theme 359(A), are visualized in red, yellow, and green links, representing the shared word pairs and the pairs unique to the ASRs and SEPs/TEMPs, respectively. Figures 8, 9, 10, and 11 show that there are

concepts that are more prevalent in the ASRs than in the SEPs and TEMPs, or that appear uniquely in the ASRs or in the SEPs and TEMPs. Because the SEPs and TEMPs documents are supposed to support the ASRs documents, the illustrations and visualizations of LLA might inform acquisition professionals about why concepts in the SEPs and TEMPs were missing from the ASRs and vice versa. FOUO word pairs and concepts about these comparisons are listed in Appendix A, which shows samples of consensus (word pairs that appear in both data sources) and gaps (word pairs that appear unique to one data source).

Comparing Figures 8 and 10, it is clear that popular themes tend to have higher correlations among data sources (ASRs, SEPs/TEMPs) while anomalous themes tend to have lower correlations between the two data sources.

	Theme Id	All Sources	_ASR	_SEP	Overlap	Correlation
2	117(E)	299	201	51	47	0.157
3	347(P)	481	346	67	68	0.141
4	395(P)	500	330	102	68	0.136
5	130(P)	590	428	89	73	0.124
6	281(P)	469	372	42	55	0.117
7	210(P)	570	400	105	65	0.114
8	298(P)	599	348	184	67	0.112
9	388(P)	508	381	73	54	0.106
10	263(P)	666	517	79	70	0.105
11	368(P)	669	472	127	70	0.105
12	330(P)	546	391	99	56	0.103
13	147(E)	234	181	29	24	0.103
14	224(E)	331	236	62	33	0.100
15	144(P)	490	350	92	48	0.098
16	270(P)	502	371	82	49	0.098
17	235(E)	431	329	60	42	0.097
18	245(E)	281	215	39	27	0.096
19	113(E)	334	245	57	32	0.096
20	310(P)	586	441	90	55	0.094
21	182(A)	197	157	22	18	0.091

Figure 8. Themes for Comparing SEPs and ASRs, Sorted According to Ascending Correlation

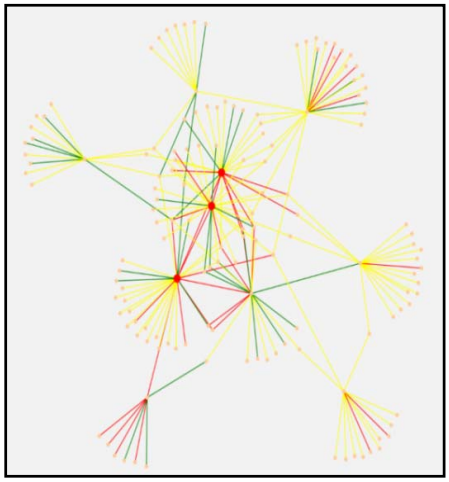


Figure 9. Detail of Word Pairs for Theme 117(E).

Note. Red Links are for shared word pairs for SEPs and ASRs. Yellow Links are for word pairs unique to ASRs, and green links for word pairs Unique to SEPs.

	Theme Id	All Sources	ASR	SEP	Overlap	Correlation
2	359(A)	149	127	18	4	0.027
3	390(A)	173	150	18	5	0.029
4	419(A)	95	73	18	4	0.042
5	267(A)	149	123	19	7	0.047
6	238(A)	170	121	41	8	0.047
7	293(A)	231	184	36	11	0.048
8	76(E)	249	208	28	13	0.052
9	408(E)	419	376	21	22	0.053
10	287(A)	223	187	24	12	0.054
11	203(E)	259	170	75	14	0.054
12	334(E)	276	218	43	15	0.054
13	135(E)	271	218	38	15	0.055
14	104(A)	196	163	22	11	0.056
15	63(E)	314	253	43	18	0.057
16	373(P)	480	403	49	28	0.058
17	372(P)	608	509	62	37	0.061
18	389(A)	155	137	8	10	0.065
19	331(E)	383	246	112	25	0.065
20	205(P)	561	420	104	37	0.066
21	127(P)	490	414	43	33	0.067

Figure 10. Themes for Comparing SEPs and ASRs, Sorted According to Descending Correlation

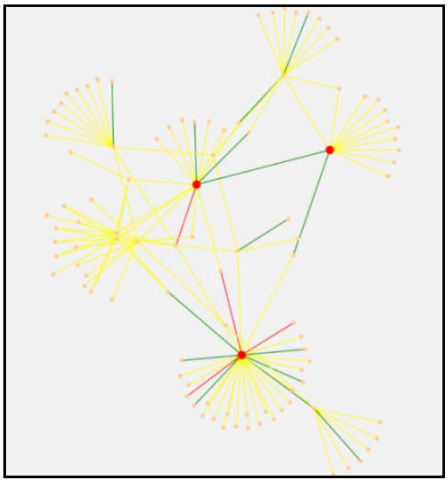


Figure 11. Detail of Word Pairs for Theme 359(A)

Note. Red links are for shared word pairs for SEPs and ASRs. Yellow links are for word pairs unique to ASRs and green Links are for word pairs unique to SEPs.

In Table 1, the highlighted cells are the ones with correlation > .06. The categories “DAESs,” “SARs,” and “SEPs” have higher overall correlations with other categories. The most highly correlated two categories are “SARs” and “DAESs” (correlation = .117). The category TEMPs has the lowest overall correlations with other categories. TEMPs and SEPs were both produced in the technical communities, yet the correlation between the two is low (.027). Appendix B shows samples of FOUO word pairs representing the gaps between SEPs and TEMPs.

When discussing the findings with the domain expert, it seems the correlation is also surprisingly low for DAESs and SARs. DAESs and SARs are similar in context and content (both relate to acquisition performance), so they would be expected to have a higher correlation. Appendix C shows a sample of FOUO word pairs representing the gaps between DAESs and SARs. Further investigations are summarized as follows:

- DAESs included more details than SARs because they are FOUO, and SARs are unclassified and publically available (e.g., http://www.dod.mil/pubs/foi/logistics_material_readiness/acq_bud_fin/SARs.html)
- Differentiate the SARs and DAESs by year and compute the correlations over time to see when the significant discrepancies (i.e., the drop in the correlation) came into the picture. The LLA correlations

between DAESs and SARs by year are listed in Table 2. The 2008 SARs is missing. The correlations dropped from 2010 to 2012.

Table 1. LLA Correlations Between Categories of Information

	APB	ASR	2366B_Cert	DAES	SARs	SEP	TEMP	TRA
APB	1.000	0.007	0.027	0.022	0.080	0.014	0.010	0.005
ASR	0.007	1.000	0.015	0.048	0.025	0.075	0.028	0.001
2366B_Cert	0.027	0.015	1.000	0.026	0.038	0.026	0.018	0.068
DAES	0.022	0.048	0.026	1.000	0.117	0.073	0.023	0.003
SARs	0.080	0.025	0.038	0.117	1.000	0.044	0.020	0.004
SEP	0.014	0.075	0.026	0.073	0.044	1.000	0.027	0.003
TEMP	0.010	0.028	0.018	0.023	0.020	0.027	1.000	0.002
TRA	0.005	0.001	0.068	0.003	0.004	0.003	0.002	1.000

Table 2. LLA Correlations Between DAESs and SARs

	LLA Correlations between DAES and SARs Reports Over Years
2004	0.21
2005	0.15
2006	0.15
2007	0.20
2009	0.17
2010	0.14
2011	0.14
2012	0.12

We correlated the DAESs or SARs over time, separately, to see if the correlation increases and decreases might have to do with the new features being introduced into the program, and, therefore, correlate to the significance of low or high changes found in LLA with the numeric metrics such as cost, schedule, funding, and performance.

Table 3 and Table 4 show correlation matrices generated by LLA for DAESs and SARs over the years. 2004's correlations for both DAESs and SARs decreased over the years (e.g., with 2005, 2006, etc.), and 2012's correlations increased over the years (e.g., with 2005, 2006, etc.). The patterns show the content gradually changed and reflected continuous innovations over the years.

Table 3. Correlation Matrix Over the Years for DAES

	DAES_2004	DAES_2005	DAES_2006	DAES_2007	DAES_2008	DAES_2009	DAES_2010	DAES_2011
DAES_2004								
DAES_2005	0.36							
DAES_2006	0.32	0.37						
DAES_2007	0.27	0.24	0.28					
DAES_2008	0.33	0.28	0.30	0.39				
DAES_2009	0.24	0.22	0.22	0.27	0.46			
DAES_2010	0.21	0.19	0.20	0.20	0.28	0.38		
DAES_2011	0.16	0.16	0.16	0.18	0.21	0.28	0.31	
DAES_2012	0.12	0.12	0.13	0.14	0.15	0.21	0.22	0.29

Table 4. Correlation Matrix Over the Years for SARs

	SARS_2004	SARS_2005	SARS_2006	SARS_2007	SARS_2009	SARS_2010	SARS_2011	SARS_2012
SARS_2004								
SARS_2005	0.53							
SARS_2006	0.52	0.54						
SARS_2007	0.46	0.45	0.51					
SARS_2009	0.46	0.44	0.50	0.50				
SARS_2010	0.27	0.24	0.26	0.27	0.28			
SARS_2011	0.29	0.25	0.27	0.27	0.27	0.46		
SARS_2012	0.24	0.22	0.23	0.23	0.24	0.39	0.60	

Figure 12 shows that estimated cost and funding for current years were taken from the DAES reports.

Cost and Funding

Cost Summary

Total Acquisition Cost and Quantity

Appropriation	BY2010 $M			BY2010 $M	TY $M		
	Initial Development APB	Current APB Development Objective/Threshold		Current Estimate	Initial Development APB	Current APB Development Objective	Current Estimate
RDT&E	3433.3	3433.3	3776.6	3391.4	3481.7	3481.7	3457.3
Flyaway	--	--	--	3391.4	--		3457.3
Recurring	--	--	--	0.0	--	--	0.0
Non Recurring	--	--	--	3391.4	--	--	3457.3
Support	--	--	--	0.0	--	--	0.0
Procurement	28369.2	28369.2	31206.1	27061.9	33720.5	33720.5	33881.4
Flyaway	28369.2	--	--	27061.9	33720.5	--	33881.4
Recurring	28090.9	--	--	26949.6	33401.8	--	33746.6
Non Recurring	278.3	--	--	112.3	318.7	--	134.8
Support	0.0	--	--	0.0	0.0	--	0.0
Other Support	0.0	--	--	--	0.0	--	--
Initial Spares	0.0	--	--	--	0.0	--	--
MILCON	208.5	208.5	229.4	202.4	236.6	236.6	236.6
Acq O&M	0.0	0.0	--	0.0	0.0	0.0	0.0
Total	32011.0	32011.0	N/A	30655.7	37438.8	37438.8	37575.3

Figure 12. Current Estimated RDT&E Cost

Table 5. Comparison of RDT&E Funding/Cost With LLA Correlations

Year	LLA correlation for DAES	LLA Correlation for SARs	RDT&E Funding/Cost in DAES ($M)	RDT&E Funding/Cost in SARs ($M)	Actual RDT&E Reported in Program Elements (PE)
2004			1314.9	1313.7	158.3
2005	0.36	0.53	1328.9	1701.9	452.6
2006	0.37	0.54	1701.9	1938.9	584
2007	0.28	0.51	1938.9	2848.6	663
2008	0.39		1211.7		309
2009	0.46	0.5	1211.7	3732.5	372
2010	0.38	0.28	3732.5	3481.7	421
2011	0.31	0.46	3481.7	3457.3	191
2012	0.29	0.6	3457.3	3387.1	297

Our observations are summarized in the following:

First, from Table 5, based on the LLA correlations for both DAES and SARs, there seem to be two periods for the program: 2004 to 2009 and 2010 to 2012. Compared to the period of 2004 to 2009, the LLA correlations in 2010 to 2012 decreased. This observation may mean that in 2010, new elements were added to the program or a new phase of the program started. Some explanation for this discontinuity might be related to a critical breach (also known as a Nunn-McCurdy breach, which is the legislation that forces the Pentagon to certify the program's fitness to continue and provides for potential congressional involvement).

Second, in Table 5, lower LLA correlations in both SARs and DAESs from previous years indicate higher RDT&E funding/cost for the current years, evidently from the following negative correlations:

- The Pearson correlation between Column 2 (LLA Correlation for DAESs) and Column 4 (RDT&E Funding/Cost in DAESs) is -.5.
- The Pearson correlation between Column 3 (LLA Correlation for SARs) and Column 5 (RDT&E Funding/Cost in SARs) is -.35.
- The Pearson correlation between Column 1 (LLA Correlation for SARs) and Column 6 (Actual Cost in PE) is -.1.

This discovery confirms that current defense acquisition practice tends to fund new elements and innovation of a program as a resource management strategy.

Finally, LLA offers a more detailed view of the difference between the 2004 DAESs and the 2010 DAESs by listing the key pairs that are associated with each theme sorted by their correlation in a descending order. Each theme can be described using a set of key words, word pairs reflecting the difference in the two DAESs. The themes on the top of the list reflect the unique themes in the 2010 DAESs as shown by green word links in the FOUO Appendix D, and the ones on the bottom of the list reflect the themes with more shared word pairs in both reports (correlation > .2) as shown the red word pair links in the FOUO Appendix D.

Conclusion

This is the first program to have undergone a relatively comprehensive LLA analysis to generate semantic networks of the acquisition artifacts among multiple categories of data. First, we found that there are many consensus or consistency areas existing in the various artifacts, yet gaps or low correlations seem to characterize the relations among these categories of artifacts, for example, between ASRs and SEPs/TEMPs, between SEPs and TEMPs, and between SARs and DAESs. Specifically, many concepts in one category are not documented in another, which could be the basis for further investigations. LLA is able to discover in detail where the gaps and inconsistencies of the data are across various communities (e.g., engineering and acquisitions communities). The semantic networks discovered using LLA, reported as themes, concepts, and word pairs, may, however, help identify the issues and offer specific and productive directions for further examination as to why there are gaps and where the initial indications of the data's consistency or discrepancy are shown as FOUO appendices of this report.

This is a major advantage of using LLA. When the correlations are in turn correlated with the cost/funding data over the years, decision-makers may then see,

in a big picture across all the programs in AVP, where data gaps exist, how and if trends over the years make sense, and how and if the acquisition strategies for these programs earn the desired return on investment (ROI) in terms of resource management and big data management.

Future Work

Much more work is needed in this area and continued in-depth analysis must be performed at the different levels of the AVP. We will continue to seek to show that LLA can be adapted to the AVP's ongoing requirements and continuous improvement of DoD data quality and decision-making. The following are the directions for the future work:

- Continue working with program management to resolve the data access issues so more program data can be extracted and analyzed automatically using the same methodology.
- Improve the dynamic interface of LLA so users can inquire about desired comparisons and program features that help them explore, link, and predict program risks.
- Explore meaningful ways to link numeric features such as various cost measures (life-cycle cost and ownership cost, among others) with features or independencies of multiple programs.
- Explore if LLA can use ACQuipedia (https://dap.dau.mil/acquipedia/Pages/Default.aspx), which serves as an online encyclopedia of common defense acquisition topics. Each topic is identified as an article; each article contains a definition, a brief narrative that provides context, and links to the most pertinent policy, guidance, tools, practices, and training that further augment understanding and expand depth. Since it contains standard terminologies for common defense acquisition topics, it might be used as supervised learning data to train LLA to improve the understanding of context-dependent meaning.

References

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. Retrieved from http://jmlr.csail.mit.edu/papers/volume3/blei03a/blei03a.pdf

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). *UCINET for Windows: Software for social network analysis*. Lexington, KY: Analytic Technologies.

Center for Computational Analysis of Social and Organizational Systems (CASOS). (2009). *AutoMap: Extract, analyze and represent relational data from texts*. Retrieved from http://www.casos.cs.cmu.edu/projects/automap/

Dahmann, J., Baldwin, K., Bergin, D., Choudhary, A., Dubon, A., & Eiserman, G. (2005). *Matrix mapping tool (MMT)*. Washington, DC: OUSD (AT&L) and Defense Systems.

Denby, E., & Gammack, J. (1999). *Modelling ignorance levels in knowledge-based decision support*. Retrieved from http://wawisr01.uwa.edu.au/1999/DenbyGammack.pdf

Dumais, S. T., Furnas, G. W., Landauer, T. K., & Deerwester, S. (1988). Using latent semantic analysis to improve information retrieval. In *Proceedings of CHI'88: Conference on Human Factors in Computing* (pp. 281–285). New York City: Association for Computing Machinery.

Freeman, L.C. (1979). Centrality in social networks I: Conceptual clarification. *Social Networks,* 1: 215-239

Feldman, R., Sanger, J. (2007). *The Text Mining Handbook*. Cambridge: Cambridge University Press.

Gallup, S. P., MacKinnon, D. J., Zhao, Y., Robey, J., & Odell, C. (2009, October 6–8). Facilitating decision making, re-use and collaboration: A knowledge management approach for system self-awareness. In *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management (IC3K).* New York City: Springer Publishing.

Girvan, M., & Newman, M. E. J. (2002, June). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *USA*, *99*(12), 7821–7826.

Hubert, L., & Schultz, J. (1976). Quadratic assignment as a general data-analysis strategy. *British Journal of Mathematical and Statistical Psychology*, *29*, 190–241.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Orgnet. (2011). Social network analysis: A brief introduction. Retrieved from http://www.orgnet.com/sna.html

Batagelj, V., Mrvar, A., & Zaveršnik, M.. (2011). Pajek - Program for large network analysis. Retrieved from http://pajek.imfm.si/doku.php?id=pajek

Quantum Intelligence (QI). (2009). Collaborative learning agents (CLA). Retrieved from http://www.quantumii.com/qi/cla.html

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge, UK: Cambridge University Press.

Toutanova, K., & Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)* (pp. 63–70). Retrieved from http://nlp.stanford.edu/manning/papers/emnlp2000.pdf

Web3D (2011). Web 3D consortium. Retrieved from http://www.web3d.org

Zhao, Y., Gallup, S. P., & MacKinnon, D. J. (2010). Towards real-time program-awareness via lexical link analysis. In *Proceedings of the Seventh Annual Acquisition Research Symposium*. Retrieved from http://acquisitionresearch.net

Zhao, Y., Gallup, S. P., & MacKinnon, D. J. (2011a). A web service implementation for large-scale automation, visualization and real-time program-awareness via lexical link analysis. In *Proceedings of the Eighth Annual Acquisition Research Program*. Retrieved from http://acquisitionresearch.net

Zhao, Y., Gallup, S. P., & MacKinnon, D. J. (2011b). Lexical link analysis for the Haiti earthquake relief operation using open data source. In *Proceedings of the 16th ICCRTS, International Command and Control, Research and Technology Symposium*. Retrieved from http://www.dodccrp.org/events/16th_iccrts_2011/papers/164.pdf

Zhao, Y., Gallup, S. P., & MacKinnon, D. J. (2011c, September). System self-awareness and related methods for improving the use and understanding of data within DoD. *Software Quality Professional*, *13*(4), 19–31. Retrieved from http://asq.org/pub/sqp/

Zhao, Y., Gallup, S. P., & MacKinnon, D. J. (2011d). *Towards real-time program awareness via lexical link analysis* (Acquisition Research Sponsored Report Series; NPS-AM-10-174). Monterey, CA: Naval Postgraduate School.

Zhao, Y., Gallup, S. P., & MacKinnon, D. J. (2012, May). Applications of lexical link analysis web service for large-scale automation, validation, discovery, visualization and real-time program-awareness. In *Proceedings of the Ninth Annual Acquisition Research Symposium*. Retrieved from http://acquisitionresearch.net

Zhao, Y., Brutzman, D. & MacKinnon, D.J. (2013, May). Improving DoD energy efficiency: Combining MMOWGLI social media brainstorming with Lexical Link Analysis to strengthen the acquisition process. In *Proceedings of the Tenth Annual Acquisition Research Symposium*. Retrieved from http://www.acquisitionresearch.net/files/FY2013/NPS-LM-13-C10P05R03-061.pdf

Zhao, Y., Gallup, S. P. & MacKinnon, D.J. (2013, May). Lexical Link Analysis application: Improving web service to acquisition visibility portal. In *Proceedings of the Tenth Annual Acquisition Research Symposium*. Retrieved from http://www.acquisitionresearch.net/files/FY2013/NPS-AM-13-C10P01R010-039.pdf

THIS PAGE INTENTIONALLY LEFT BLANK

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL
555 DYER ROAD, INGERSOLL HALL
MONTEREY, CA 93943

www.acquisitionresearch.net